KU LEUVEN

HOGESCHOOL-UNIVERSITEIT BRUSSEL

Probabilistic latent feature analysis of two-way frequency data with the plfm package

Michel Meulders

KU Leuven @ HU Brussel

Outline

- Type of data and substantive questions
- Probabilistic latent feature model (PLFM)
- Comparison with correspondence analysis
- Description of the plfm package
- Examples

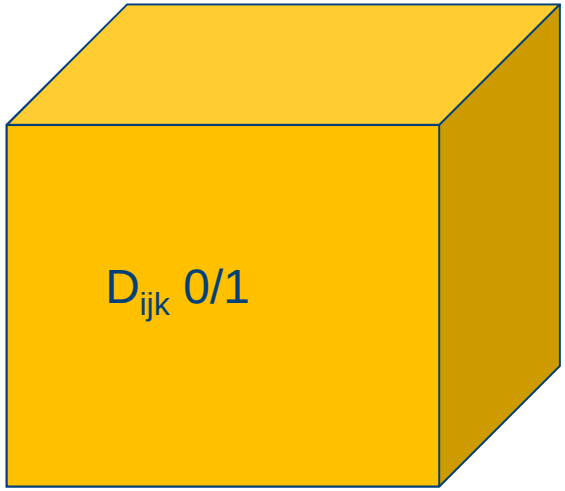
Type of data and substantive problems

- Three-way three-mode binary data: rater judgements about object x attribute associations
- Of interest in different substantive domains
  - Perceptual mapping of products: product x attribute x consumer
  - Psychiatric diagnosis: patient x symptom x clinician
  - Personality psychology: situation x behavior x person
  - Social network analysis: actor x actor x rater
  - Emotion perception: facial expression x emotion x rater
  - …

Type of data and substantive problems

- Goal of the analysis: build parsimonious models to explain observed object-attribute associations
  - Nonspatial “classification based” techniques: derive a classification of objects, attributes, raters
  - Spatial “dimension-reduction” techniques: derive a low-dimensional spatial representation of objects, attributes, raters
- Starting point: analysis of two-way frequency data obtained by aggregating three-way data across raters
- However: rater differences are often often of key-interest

Three-way three-mode binary data



Rater i
(i=1,..,I)

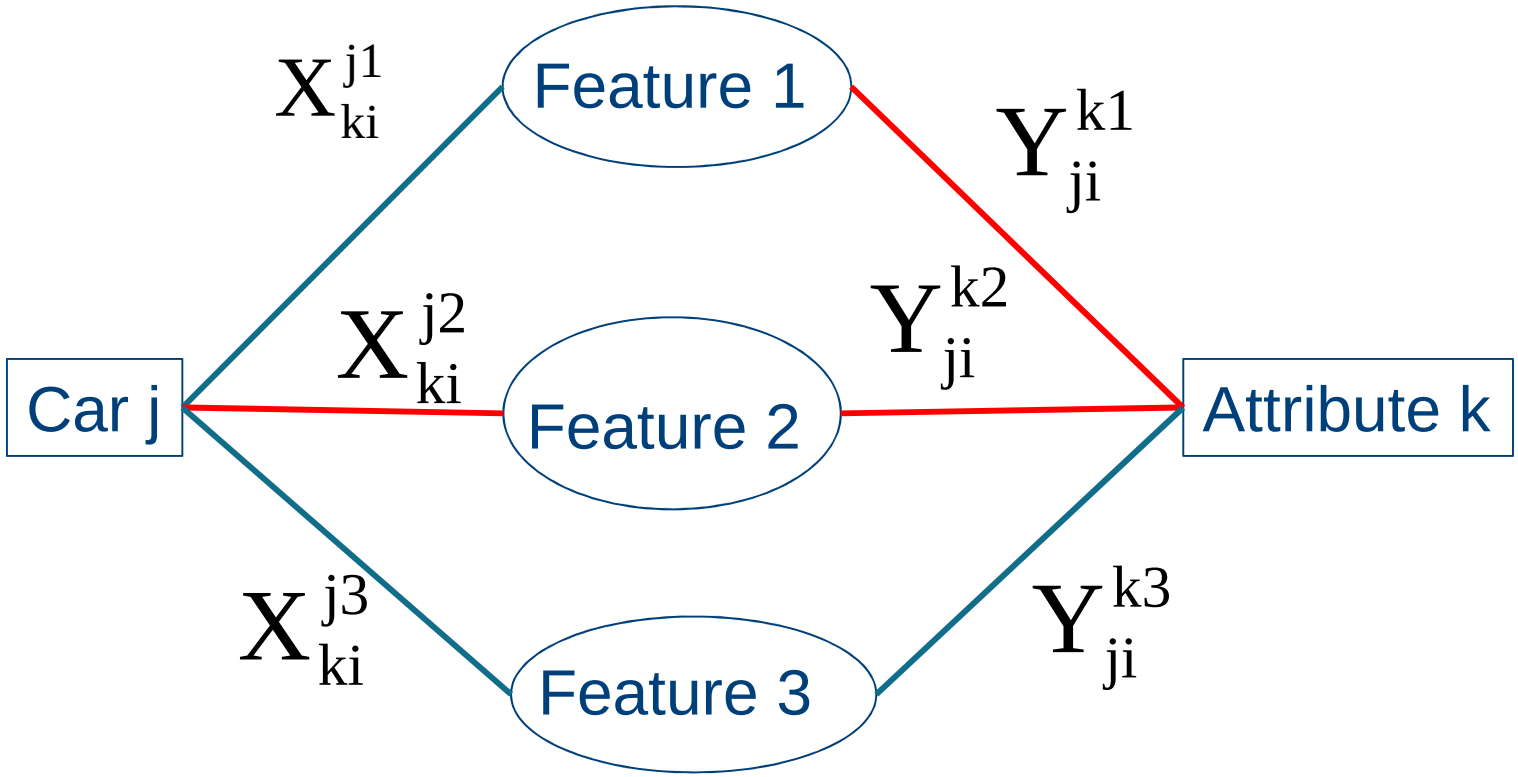
Car j
(j=1,..,J)

Attribute k
(k=1,..,K)

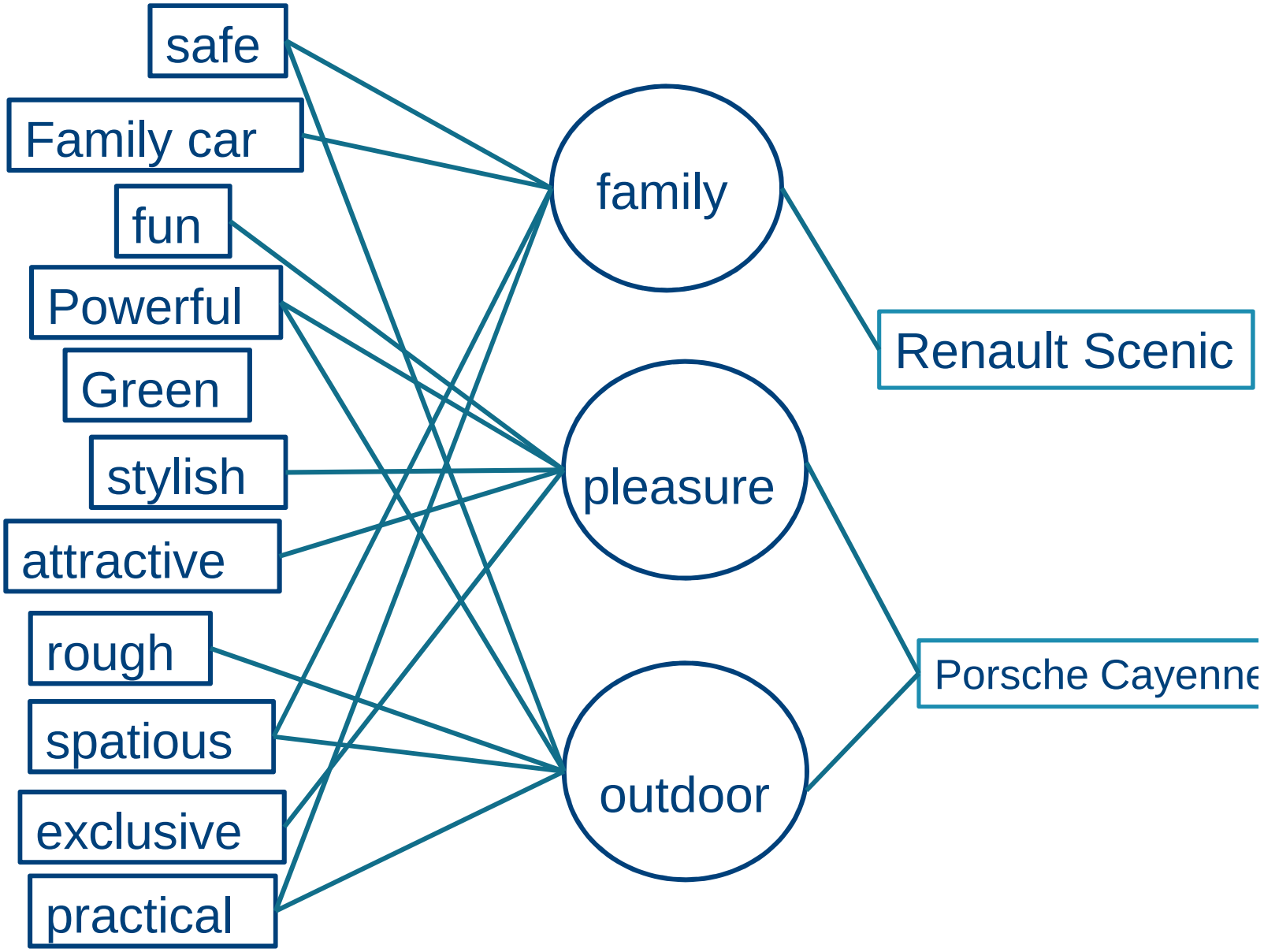
D_{ijk}=1 if car j has attribute k according to rater i
D_{ijk}=0 otherwise

Probabilistic latent feature models

Explain observed associations based on binary latent features:



$$D_{ijk} = 1 \Leftrightarrow \exists\, f : X_{ki}^{jf} = Y_{ji}^{kf} = 1$$ (disjunctive rule)


safe
Family car
fun
Powerful
Green
stylish
attractive
rough
spatious
exclusive
practical
family
pleasure
outdoor
Renault Scenic
Porsche Cayenne

Probabilistic latent feature models

- It is assumed that

$$X_{ki}^{jf} \sim \text{Bern}(\sigma_{jf})$$

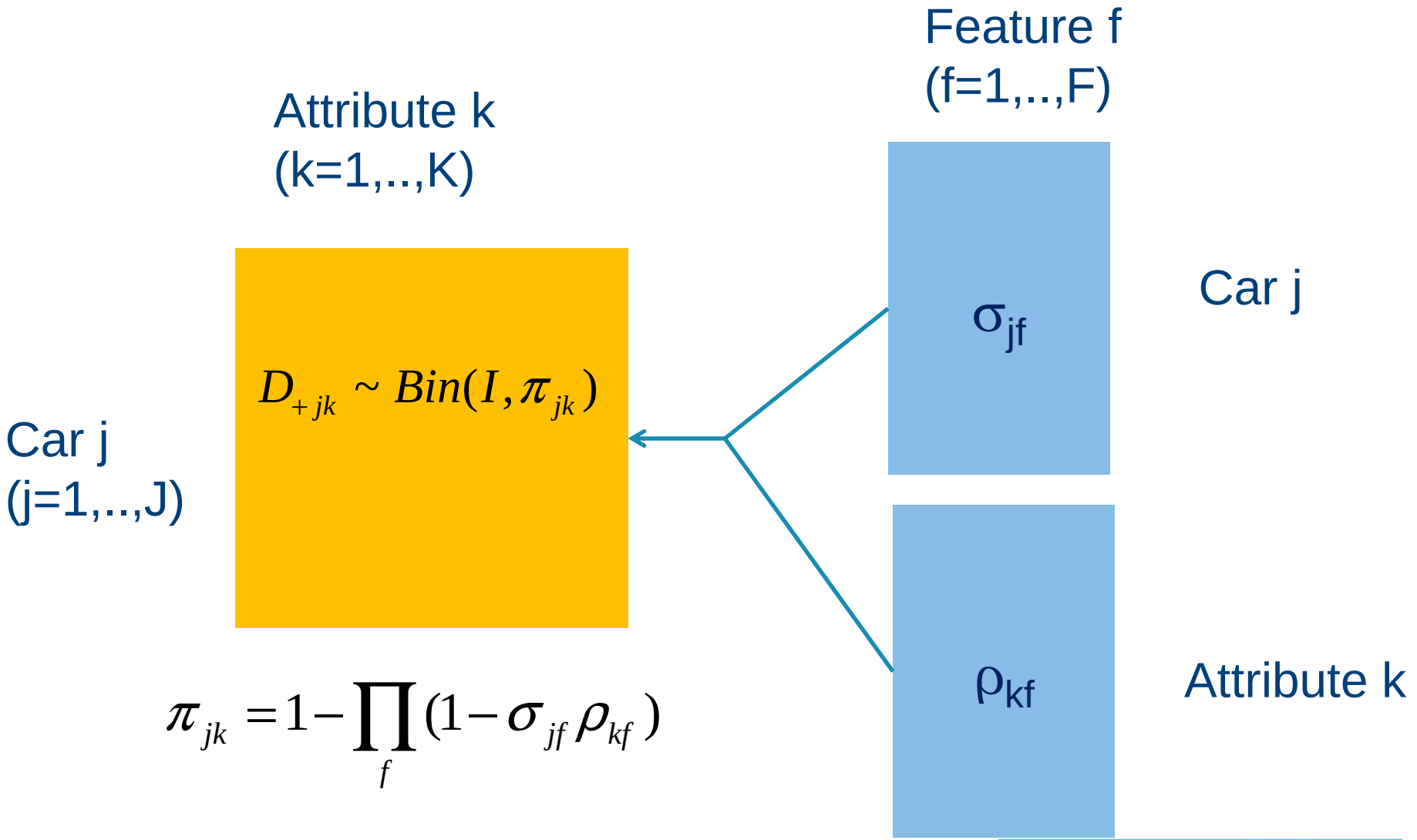
$$Y_{ji}^{kf} \sim \text{Bern}(\rho_{kf})$$

$$D_{ijk} = 1 \Leftrightarrow \exists f : X_{ki}^{jf} = Y_{ji}^{kf} = 1$$

- From the distribution of X, Y and the mapping rule D=f(X,Y) it follows that

$$\pi_{jk} = P(D_{ijk} = 1) = 1 - \prod_{f} (1 - \sigma_{jf} \rho_{kf})$$

Probabilistic latent feature models



Correspondence analysis (CA)

$$P_{jk} = \frac{D_{+jk}}{D_{+++}}$$

Attribute k
(k=1,..,K)

Car j
(j=1,..,J)

$$S_{jk} = \frac{P_{jk} - P_{j+}P_{+k}}{\sqrt{P_{j+}P_{+k}}}$$

$$\approx \hat{S}_{jk} = \sum_{q=1}^{Q} \delta_q u_{jq} v_{kq}$$

SVD

Principal dimension q
(q=1,..,Q)

u_{jq} Car j

$Diag(\delta_q)$ Dimension q

v_{kq} Attribute k

Theoretical comparison PLFM and CA

Aspect	PLFM	CA
Representation of objects and attributes	Feature-based (nonspatial)	dimension-based
Dependence derived features/dimensions	Correlated features	Uncorrelated principal components
Type of model	Non-compensatory (disjunctive, conjunctive,..)	compensatory

Probabilistic latent feature analysis with plfm package

- The plfm package contains functions for probabilistic latent feature analysis.
- The plfm() function is applied to an object x attribute frequency matrix and yields for a specific PLFM
  - Point estimates of object and attribute parameters and asymptotic standard errors
  - Model selection criteria (AIC, BIC)
  - Descriptive goodness-of-fit measures (i.e. correlation between observed and expected frequenies)
  - Statistical test of absolute goodness-of-fit (i.e. Pearson chi-square)

Probabilistic latent feature analysis with plfm package

- The function stepplfm() can be used to estimate a series of disjunctive and/or conjunctive models that assume minF to maxF latent features
- A plot() function can be used to visualize the fit of models (AIC, BIC, VAF,…) as a function of the number of latent features
- Summary() and print() functions are available to generate a report about the fitted models

Probabilistic latent feature analysis with plfm package

- The bayesplfm() function can be used to compute a sample of the observed posterior distribution. This function yields
  - Draws of the posterior distribution for each parameter
  - Point estimates (i.e. posterior mean) and 95% posterior intervals for each parameter
  - Assessment of convergence

Example: product perception of car models

- Data on 78 raters who judge for all pairs of 14 cars and 27 attributes whether a car has an attribute

Cars

Volkswagen Golf	Opel Corsa	Nissan Qashgai	Toyota Prius
BMW X5	Volvo V50	Renault Espace	Citroen C4 Picasso
Ford Focus Cmax	Mercedes C-class	Fiat 500	Audi A4
Mini Cooper	Mazda MX5		

Attributes

Economical	Agile	Environmentally friendly
Reliable	Practical	Family Oriented
Versatile	Good price-quality ratio	Luxurious
Safe	Sporty	Attractive
Comfortable	Powerful	Status symbol
Technically advanced	Sustainable	Original
Nice design	Value for the money	High trade-in value
Exclusive	Popular	Outdoor
Green	City focus	Workmanship

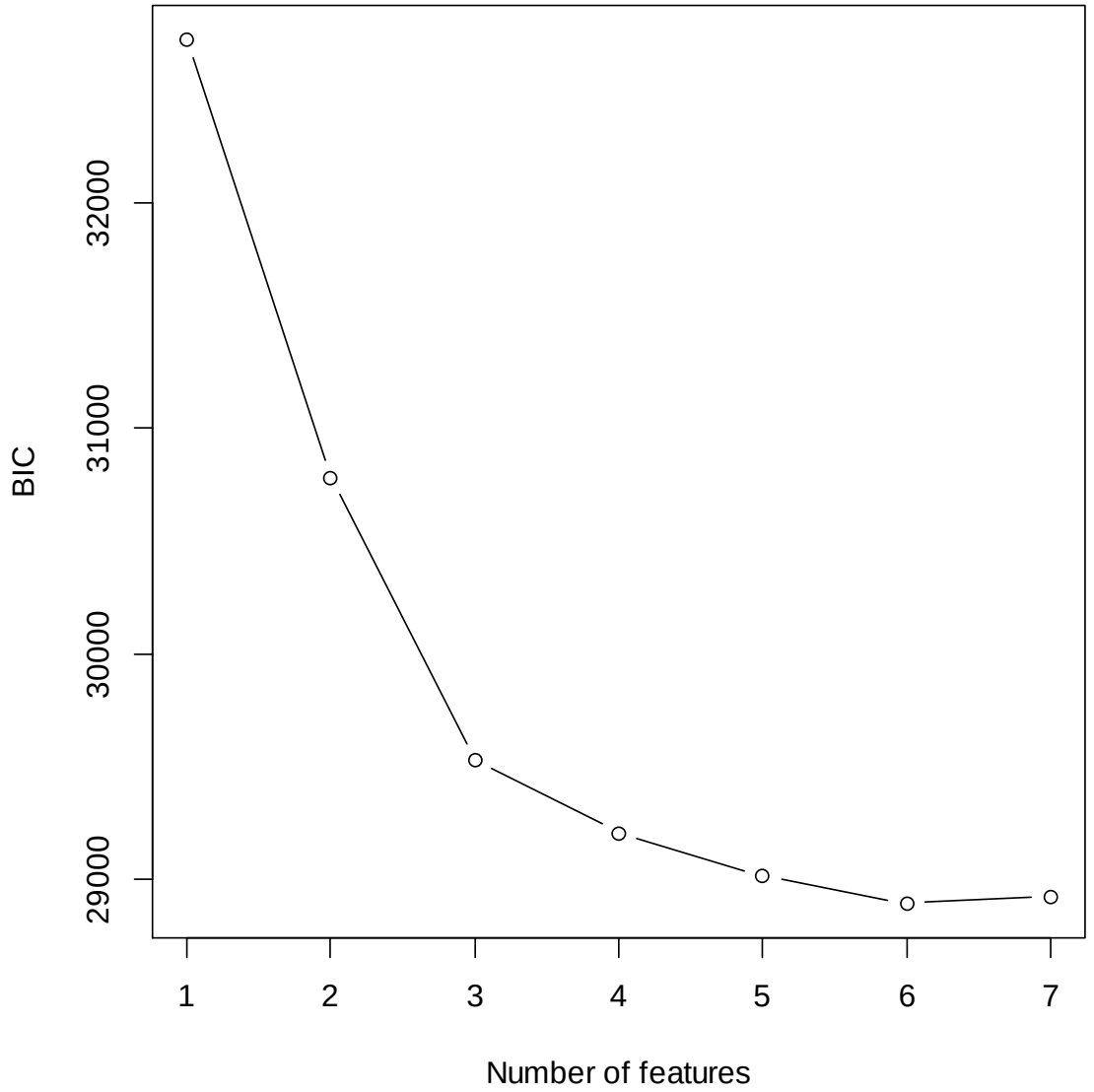
Example: product perception of car models

```
## load the package
R> library("plfm")
## load the car data
R> data("car")
## use stepplfm() to locate posterior modes of
##disjunctive models with 1 up to 7 features,
##estimation of each model is based on 20 runs
R> car.lst<-stepplfm(freq1=car$freq1,freqtot=78,
+ maprule="disj",minF=1,maxF=7,M=20)
```

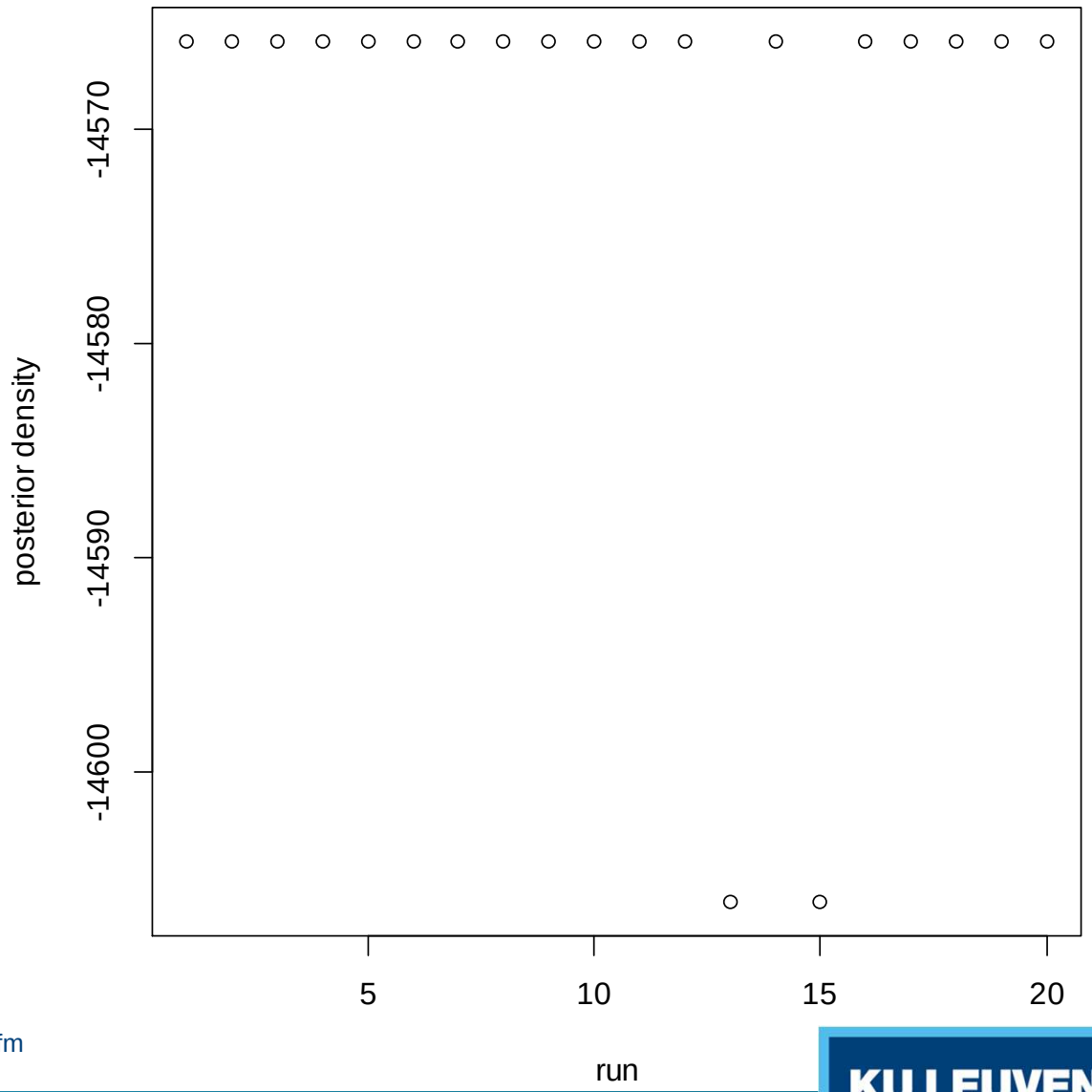
Example: product perception of car models

```
## print output
R> Car.lst
INFORMATION CRITERIA:
     LogLik LogPost Deviance   AIC   BIC
F=1 -16271  -16332    32542 32624 32721
F=2 -15210  -15375    30420 30584 30777
F=3 -14496  -14792    28991 29237 29527
F=4 -14245  -14661    28489 28817 29204
F=5 -14060  -14591    28121 28531 29014
F=6 -13910  -14566    27820 28312 28892
F=7 -13835  -14605    27670 28244 28921
```

```
R> plot(car.lst,which="BIC")
```



```
plot(car.lst[[6]]$logpost.runs,xlab="run",ylab="posterior density",
main="posterior density runs")
```



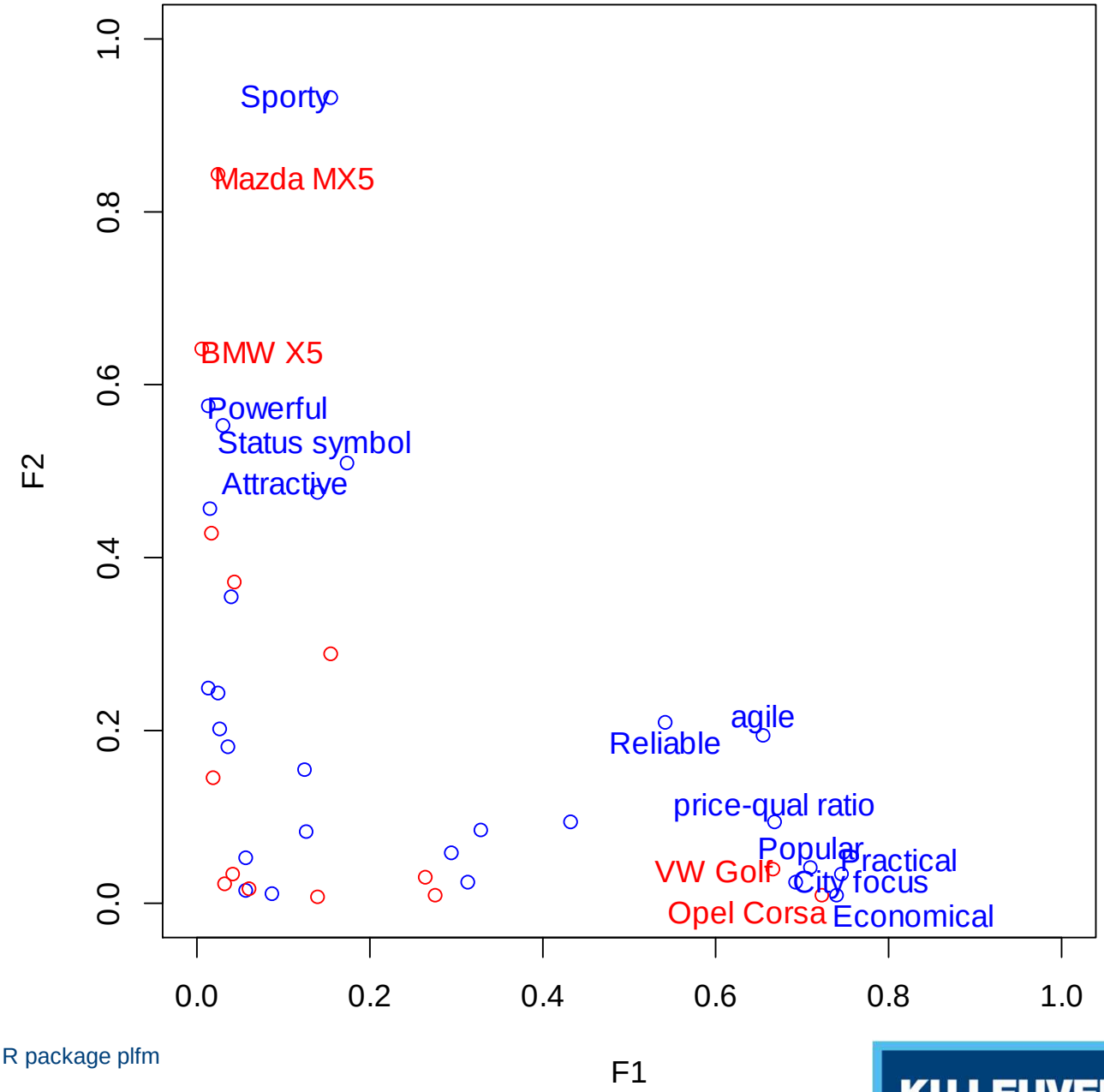
Example: product perception of car models

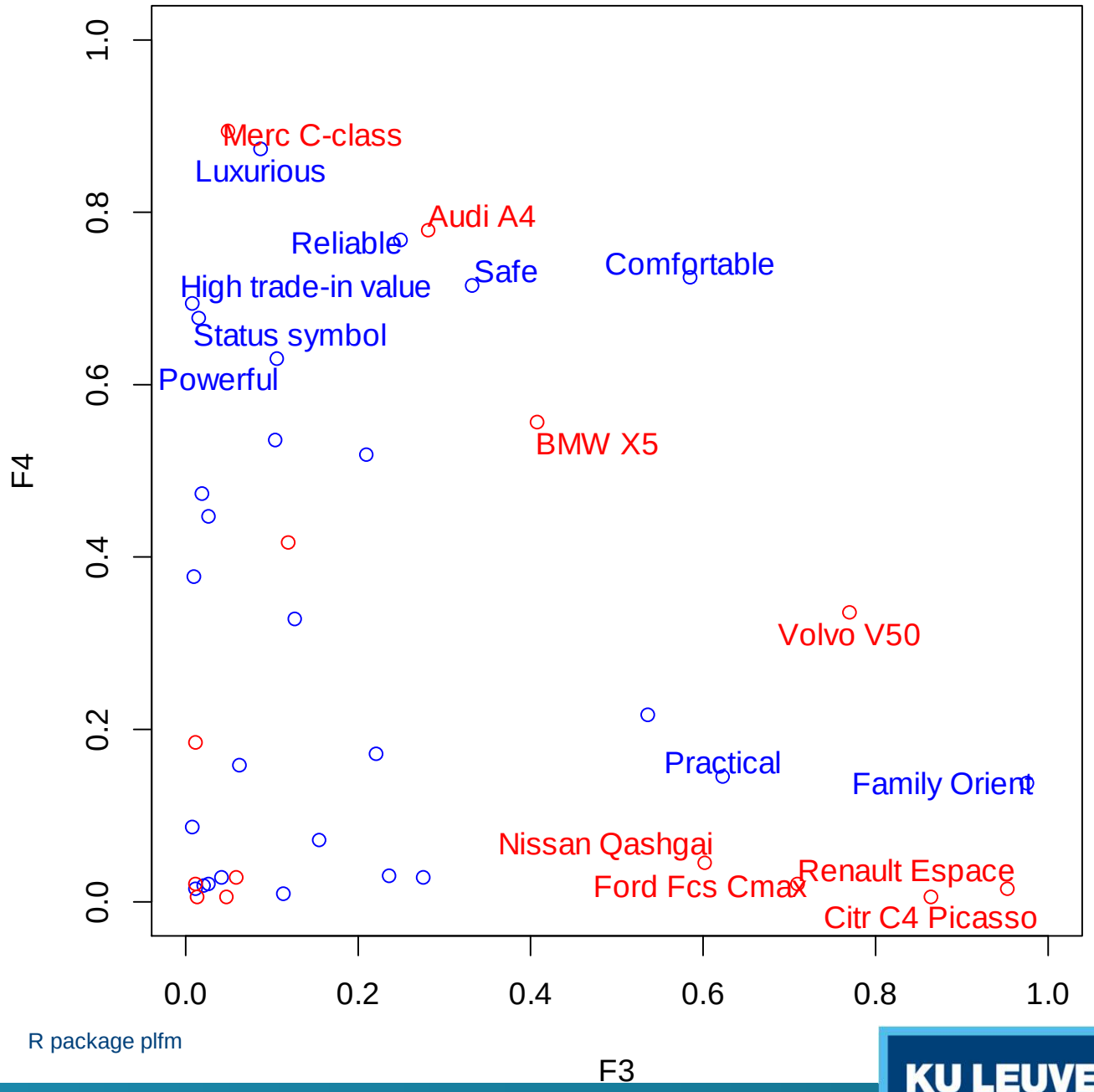
```
PEARSON CHI SQUARE TEST OBJECT X ATTRIBUTE TABLE:
    Chisquare  df p-value
F=1  5340.581 337       0
F=2  3228.952 296       0
F=3  1785.729 255       0
F=4  1267.097 214       0
F=5   858.555 173       0
F=6   570.581 132       0
F=7   422.934  91       0
```

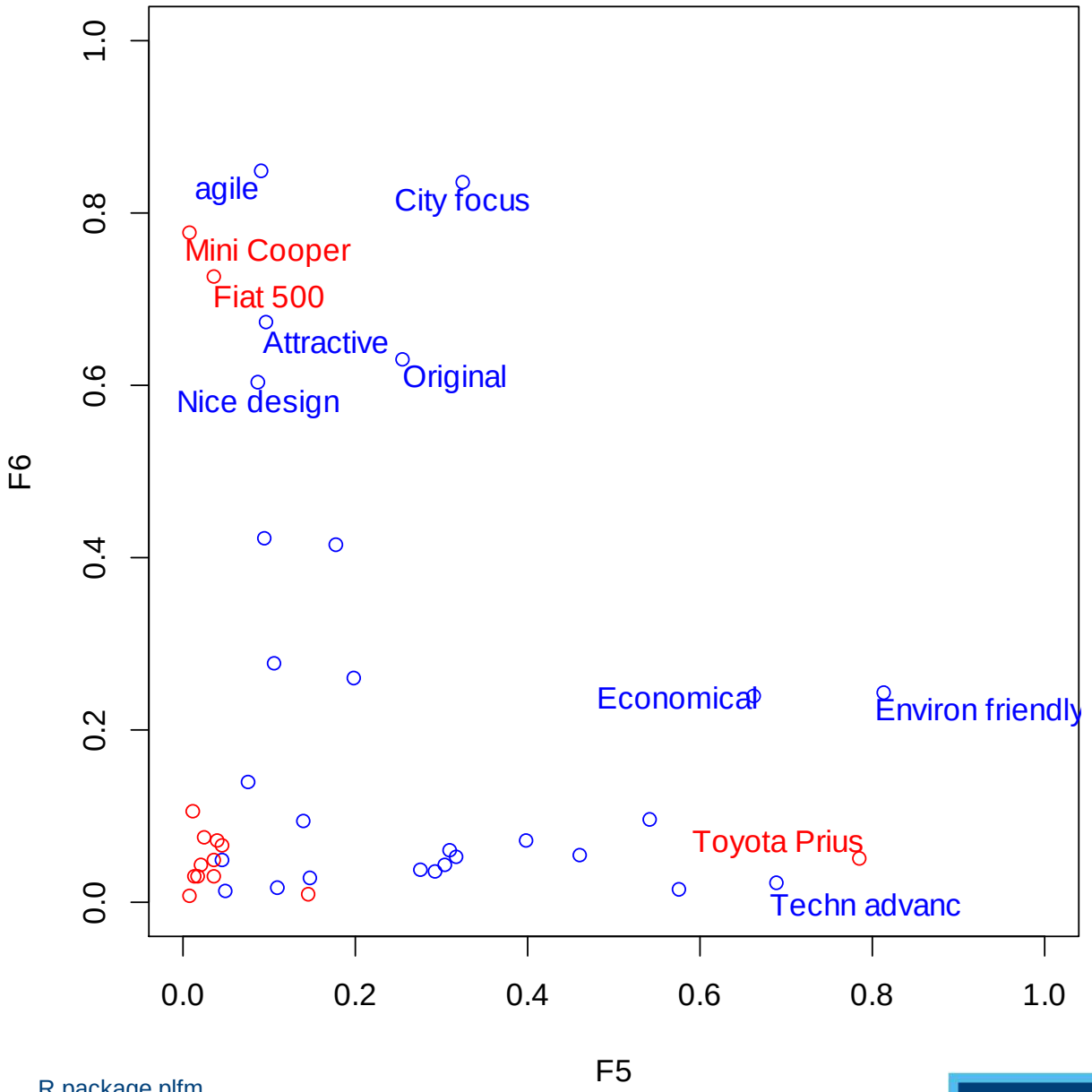
Example : product perception of car models

DESCRIPTIVE FIT OBJECT X ATTRIBUTE TABLE:

	Correlation	VAF
F=1	0.516	0.267
F=2	0.742	0.551
F=3	0.871	0.759
F=4	0.911	0.830
F=5	0.940	0.883
F=6	0.960	0.922
F=7	0.973	0.947


Sporty
Mazda MX5
BMW X5
Powerful
Status symbol
Attractive
agile
Reliable
price-qual ratio
Popular
Practical
VW Golf
City focus
Opel Corsa
Economical
F1
F2
0.0
0.2
0.4
0.6
0.8
1.0

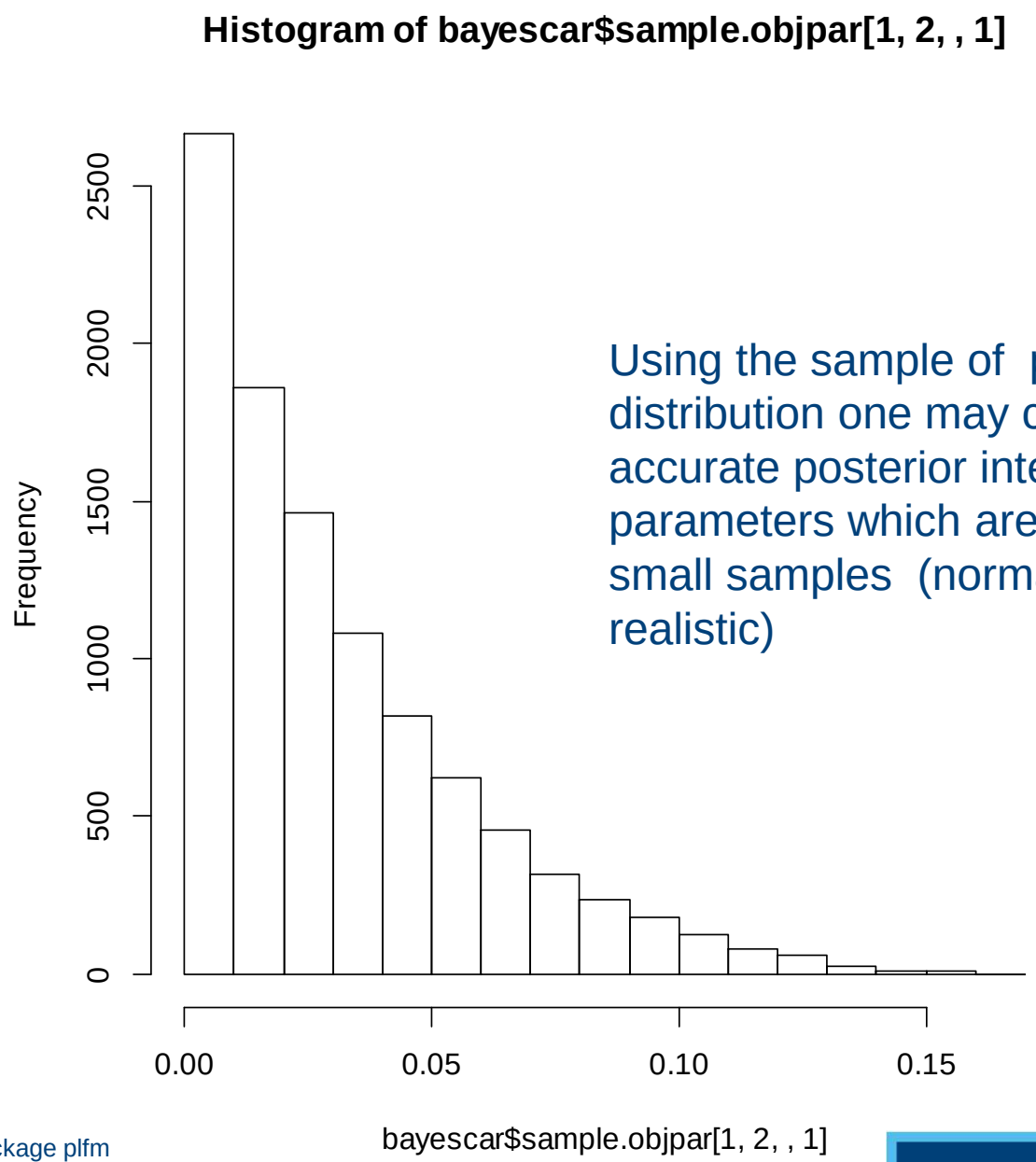

Merc C-class
Luxurious
Audi A4
Reliable
Safe
Comfortable
High trade-in value
Status symbol
Powerful
BMW X5
Volvo V50
Practical
Family Orient
Nissan Qashgai
Ford Fcs Cmax
Renault Espace
Citr C4 Picasso
F4
F3
0.0
0.2
0.4
0.6
0.8
1.0


agile
City focus
Mini Cooper
Fiat 500
Attractive
Original
Nice design
Economical
Environ friendly
Toyota Prius
Techn advanc
F6
F5
0.0
0.2
0.4
0.6
0.8
1.0

Bayesian probabilistic feature analysis

```
## compute as sample of the posterior distribution for
the disjunctive 6-feature using the best posterior mode
as a starting point
R> bayescar<-bayesplfm(maprule="disj",
+ freq1=car$freq1,freqtot=car$freqtot, F=6,
+ maxNiter=10000, Nburnin=0, Nstep=1000, Nchains=2,
+ start.bayes="fitted.plfm", fitted.plfm=car.lst[[6]])


## compute correlation between posterior mean and
##posterior mode
R> cor(c(car.lst[[6]]$attpar),c(bayescar$pmean.attpar))
[1] 0.9958402
R> cor(c(car.lst[[6]]$objpar),c(bayescar$pmean.objpar))
[1] 0.9959103
```



Using the sample of posterior distribution one may compute more accurate posterior intervals for the parameters which are also valid for small samples (normality is not always realistic)

summary

- Probabilistic latent feature models can be used to explain two-way object x attribute frequency data on the basis of a limited number of binary latent features
- The model provides a fuzzy overlapping clustering of both the objects and the attributes
- Analysis can be done with the R package plfm which provides
  - Points estimates, standard errors
  - Model selection criteria
  - Statistical and descriptive measures of goodness-of-fit
- Reference: Meulders, M. (2013). An R package for probabilistic latent feature analysis of two-way two-mode frequencies. *Journal of Statistical Software, 54*(14), 1-29. http://www.jstatsoft.org/v54/i14